GSEA Documentation

Description: Gene Set Enrichment Analysis
Author: Aravind Subramanian, Pablo Tamayo (gp-help@broad.mit.edu)

Summary:
Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes). Detailed information is available at http://www.broad.mit.edu/gsea.

Known Issues:
Input expression datasets with a '-' in their file names causes GSEA to error.

Interpretation of cls files differs from other modules. For example, in the cls file below

21 2 1
resistant sensitive
1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0

most GenePattern modules would interpret the first 11 samples to be sensitive and the remaining 10 to be resistant. However, GSEA assigns resistant to the first 11 samples and sensitive to the rest. This is because GSEA assigns the first name in the second line to the first symbol found on the third line.

Parameters:

Name	Description
expression dataset	Dataset file - .res, .gct
gene sets database	Gene sets database from GSEA website.
gene sets database file	Gene sets database - .gmt, .gmx, .grp. Upload a gene set if your gene set is not listed as a choice for the gene sets database parameter.
number of permutations	Number of permutations to perform
phenotype labels	Cls file - .cls, must be binary
collapse dataset	Select True to have GSEA collapse each probe set in the expression dataset into a single vector for the gene, which gets identified by its gene symbol.
permutation type	Type of permutations to perform
chip platform	Chip file from GSEA website.
chip platform file	Chip to use. Upload a chip file if your chip is not listed as a choice for the chip platform parameter.
scoring scheme	The statistic used to score hits (gene set members) and misses (non-members)
metric for ranking genes	Class separation metric - gene markers are ranked using this metric to produce the gene list

gene list ordering mode	Direction in which the gene list should be ordered
gene list sorting mode	Mode in which scores from the gene list should be considered
max gene set size	Gene sets larger than this are excluded from the analysis
min gene set size	Gene sets smaller than this are excluded from the analysis
collapsing mode for probe sets with more than one match	Collapsing mode
normalization mode	Normalization mode
randomization mode	Type of phenotype randomization (does not apply to gene set permutations)
omit features with no symbol match	If there is no known gene symbol match for a probe set, omit it from the collapsed dataset
make detailed gene set report	Create detailed gene set report (heat map, mountain plot, etc.) for every enriched gene set
median for class metrics	Use the median of each class instead of the mean for the class separation metrics
number of markers	Number of markers
plot graphs for the top sets of each phenotype	Plot GSEA mountain and related plots for the top sets of each phenotype
random seed	Seed to use for randomization
save random ranked lists	Whether to save random ranked lists (might be very large)
output file name	Name of the output file

Return Value:

1. zip file containing the results

References:

- Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. October 25, 2005, vol. 102, no. 43, 15545-15550